Wax On, Wax Off: Relearning How to Teach Data Analysis. An Interactive Visual Approach

FODAVA Technical Report

Scotland C. Leman and Leanna House

Department of Statistics

Virginia Tech, Blacksburg, VA 24061

In the 1984 movie, "The Karate Kid", a young teenager named Daniel LaRusso triumphs in a local martial arts tournament while coached by his apartment's handyman, Mr. Miyagi. Mr. Miyagi used a unique coaching style that developed Daniel's fundamental skills in martial arts *before* teaching their context and application. Initially, the way by which Mr. Miyagi coached confused Daniel. Time after time, Daniel went to Mr. Miyagi's home expecting to learn standard "karate moves" (e.g., how to punch), but, ended up doing repetitive household chores, such as, painting fences and waxing cars. The only instructions Mr. Miyagi would offer pertained to the chores themselves: "left hand, right hand," "up, down," "wax on, wax off," or "breath in, breath out." Although Mr. Miyagi knew his motives and felt gratified in his teachings, Daniel was frustrated. He was desperate to learn Karate, yet, to him, he was only doing tedious chores. Ultimately, Daniel lost his cool and said,

Daniel: "I'm being your *expletive* slave... we made a deal... you're supposed to teach, and I'm supposed to learn... I haven't learned a *expletive* thing".

Miyagi: "You learn plenty".

Daniel: "I'm going home, man!".

Subsequently, Mr. Miyagi showed by example how the chores translated to karate. Each chore had a purpose and taught Daniel how to move his arms (up and down and left to right) to defend himself. From that point onward, Daniel continuously built upon his karate abilities and trusted Mr. Miyagi.

Today, we teach Introductory Statistics, or more broadly, Introductory Data Analytics (DA), in a way that is similar to Mr. Miyagi's coaching style. We teach several quantitative methods using small contrived problems that, in many cases, lack relevance to the real-world. However, unlike Daniel, many students seem to miss having an "a-ha" moment when they realize how to connect classroom concepts for real-world applications. Current classes focus mainly on the quantitative aspects of DA and fail to provide opportunities for students to grasp context and think critically. For example, to apply DA thoughtfully students need critical thinking skills to compartmentalize large problems into manageable pieces, formulate and evaluate solutions with both quantitative and qualitative rigor, make judgements that assimilate current information with new, and reflect upon their judgements. Furthermore, fundamental to the application of these skills is the ability to be creative and apply current knowledge and/or analytical tools in ways that might not have been considered previously. Yet, in current DA classes, students only have opportunities to practice creative, critical thinking *after* they have mastered quantitative theory and methods.

As professors of introductory DA classes, we can do better than Mr. Miyagi. Rather than focusing only on mathematical concepts in class, we can integrate them with critical thinking so that students develop all of the skills necessary to learn from data. Interactive data visualizations (IDVs) can help with the integration. Data visualizations that adjust to

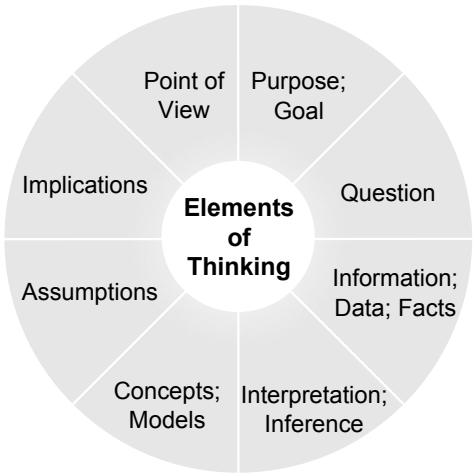


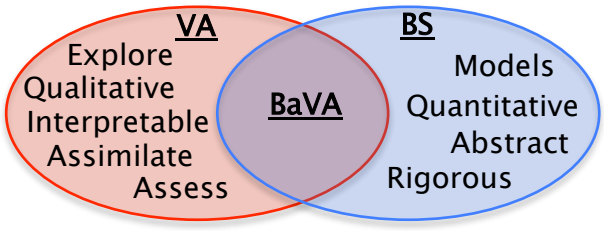
Figure 1: The "Eight Elements of Thought" (EoT) by Elder and Paul [2010].

student investigations can inspire students to make conjectures and assess them without necessarily mastering complex mathematical methods. Thus, students can apply critical thinking, as defined by the "Eight Elements of Thought" (EoT) in Figure 1, to real-world problems before, during, and after the students develop quantitative DA tools.

In this paper, we show how one may teach critical thinking with multi-dimensional scaling (MDS) and Weighted Multi-dimensional scaling (WMDS) [Kruskal and Wish, 1978] using IDVs. The IDVs that we present are based on a methodology that we call Bayesian Visual Analytics (BaVA) [House et al., 2010]. We explain BaVA in the next section using an simulated example.

1 IDV with BaVA

BaVA combines methods in Visual Analytics with Bayesian statistics so that experts may guide rigorous, quantitative data analyses via intuition and interactions (e.g., dragging, filtering, highlighting observations) with data visualizations. Mechanistically, BaVA is an iterative process that, when wrapped within interactive software, creates, adjusts, and remakes two-dimensional displays of high-



Visual Analytics (VA): "The science of analytical reasoning facilitated by interactive visual interfaces" [Thomas and Cook, 2005]. VA fosters intuitive data explorations so that experts may assess data, assimilate the information in data, and make qualitative judgements.

Bayesian Statistics (BS): A statistical paradigm that enables inference for unknowns $\boldsymbol{\theta}$ given data $\boldsymbol{d}$ and summaries of a posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{d})$. Such formulations are mathematically rigorous, abstract models that characterize datasets quantitatively.

dimensional data. The process is shown in Figure 2 and iterates as follows: 1) model data $\boldsymbol{d}$ conditional on unknowns $\boldsymbol{\theta}$ using Bayesian or data mining methods, 2) estimate and display estimates of $\boldsymbol{\theta}$ in a relevant display v, 3) prompt experts to assess and adjust the display if known or hypothesized structure in the data is missing (e.g., users may drag or highlight observations), 4) parameterize adjustments, and 5) update the original model (from step 1) so that steps 2-5 may repeat. The novelty of BaVA is that display adjustments in step 3 are considered to be reliable, expert *cognitive feedback* $f^{(c)}$ or additional data that is worth incorporating within display-generating models. Thus, the BaVA machinery interprets, quantifies, and expresses expert feedback in parametric form $f^{(p)}$ to update step 1) and repeat the process.

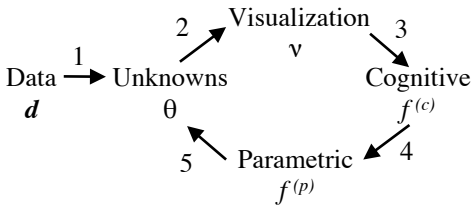


Figure 2: The BaVA process.

To exemplify BaVA, we explore a simulated dataset from House et al. [2010] that includes three groups (red, blue, and green) of three clusters, as shown in Figure 3a. Each group clusters observations based on a combination of the dimensions $\{x, y, z\}$. For example, these data could represent three genera of flowers (Rosa, Iris, and Tulipa) that each have measurements from three different species. The species may cluster based on combinations of x=*pedal width*, y=*pedal-length*, and z=*sepal-length*. To assess the clusters and uncover the dimension combinations visually (i.e., in two dimensions), we consider Principal Component Analysis (PCA) [Jolliffe, 2002]. PCA is a common DA approach that projects complex datasets to a preferred number (e.g., two) of dimensions. Each projected dimension is a "principal component."

Figure 3b displays the top two principal components (denoted $\{r_1, r_2\}$) for the data shown in Figure 3a. One benefit of data displays, is that we do not necessarily need to know how to create them- to interpret them. In this case, PCA displays the data spatially in Figure 3b so that observations that are similar and different appear close and far from each other,

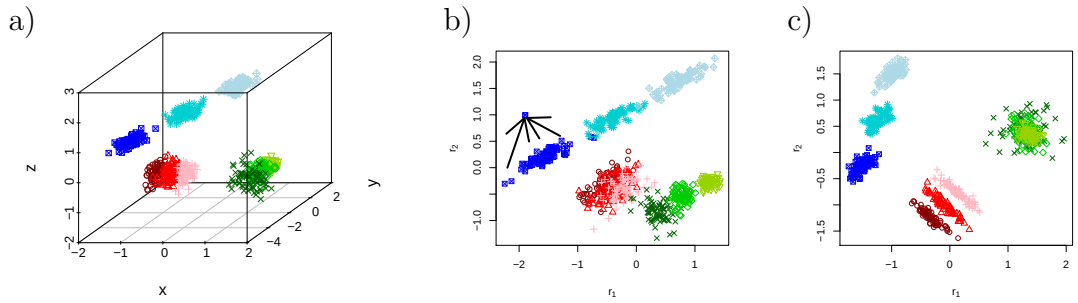


Figure 3: Figure a shows a three-dimensional simulated dataset that includes three groups (red, blue, and green) of three clusters (marked by dark, medium, and light shades of varying colors). Figure b is a PCA projection of the dataset. Notice that we cannot see some of the clusters. To express feedback about the display, two dark blue points are dragged together (as denoted by arrows in Figure b). Based on this feedback, the visualization is reconfigured using BaVA methods. Now, blue and red clusters are clear. To see the green clusters, the BaVA process can repeat.

respectively. Notice that clusters in the blue group overlap slightly. To separate the clusters and assess the dimensions (i.e., combinations $\{x, y, z\}$) that structure the blue group, we drag two dark blue points together (as depicted by arrows in Figure 3b). This is one way of expressing cognitively that we want the dimensions for which the dark blue points are similar to be represented clearly in the display. BaVA parameterizes this cognitive expression and updates the display accordingly.

Figure 3c provides the updated view of the data. Notice that we now see clusters in both the blue and red groups. By *only* moving two blue observations, we learned that the blue *and* red groups use comparable combinations of $\{x, y, z\}$ to cluster. Additionally, we learned this without an understanding of PCA.

In effect, BaVA provides a computational mediator between data characterizing models and domain experts. The experts learn from the summaries provided by the models and the models change in response to visual feedback offered by the experts. As a computational mediator, BaVA and its non-probabilistic forms serve as perfect tools to motivate, teach, and solidify methods in data analytics (DA). In the next section, we provide an example using a well known analytical approach called Multi-dimensional scaling (MDS).

2 Critical Thinking with IDV and MDS

In a course that relies on IDV to emphasize both critical thinking and DA, the focus shifts from DA methodology to solving real-world problems based on both data and personal judgement. For example, MDS, is a data visualization scheme that seeks to find a low-dimensional (e.g., two-dimensional) representation of data, e.g., a map, that portrays how the data spread in the high-dimensional space. The map results from minimizing a stress function that, to some, is hard to conceptualize. Typical approaches for teaching MDS rely, first, on explaining how to minimize the abstract stress function and, second, observing the results in a data display. Based on such approaches, students tend to memorize the MDS procedure and do not develop a comprehensive understanding of MDS. In turn, students may fail to grasp limitations and/or extensions of MDS that may apply to problems that differ slightly from classroom exercises.

On the other hand, IDVs that are based on relevant case studies motivate and enable students to draw on what they know and build an understanding of both the methods and applications of DA. In particular, we recommend teaching MDS by presenting an open-ended, real-world case study and progressing through four phases, I) Assess and Explore, II) Methods, III) Implement, and IV) Reflect. During these phases, the students use IDVs, to assess the case study, learn a DA technique (e.g., MDS, PCA), implement a technique computationally, and reflect upon results and implications.

We define these phases so that they correlate strongly with the "Elements of Thought" (EoT), as shown in Figure 1b and developed by Elder and Paul [2010]. The EoT decomposes the process of critical thinking into tangible components that, to us, captures both the quantitative and qualitative aspects of problem-solving. Thus, we deliberately focus the objectives of each phase to correspond to one or more elements of EoT so that all elements are covered by the conclusion of phase four. In the next section, we provide an example how

to teach critical thinking jointly with MDS.

3 Example

Here, we describe how to use four phases for teaching MDS using an example from Endert et al. [2011]. At the end of each section, we high-light which elements of EoT are covered within each phase. Also, Table 1 summarizes the phases by bullet points the phase objectives, categorizes the objectives as either quantitative and qualitative aspects of problem-solving, and states which elements of EoT are covered. Before we begin, we motivate the phases with a case study.

> *Case Study: The U.S. census bureau attempts to survey every individual living within the United States in order to better represent its individuals, and construct economic, health, and educational policies. We have access to a subset of the 1990 census [UCI, 1990] that includes 2.5 million observations and $p = 68$ features (i.e., variables) including wealth, education, marital status, employment status, occupation, family details, driving patterns, etc. The U.S. President (in 1992) would like to implement policy that will help those with low socio-economic status. What would you (the students) recommend? Use census data to support the recommendations.*

3.1 Phase I. Assess and Explore the Data

The way by which the case study is phrased suggests that there are multiple recommendations for the President. Thus, Phase I requires that the students 1) state in their words the goal of their endeavors, 2) hypothesize what they will learn from the data, and 3) explore the data. For the exploration, the students may look directly at an excel file that contains the data, use quantitative methods they currently know to summarize the data, and assess the data visually using BaVA software.

Figure 4a plots an initial MDS BaVA display of a random sample ($n = 3000$) from the census dataset. During Phase I, we do not explain the quantitative BaVA method used to

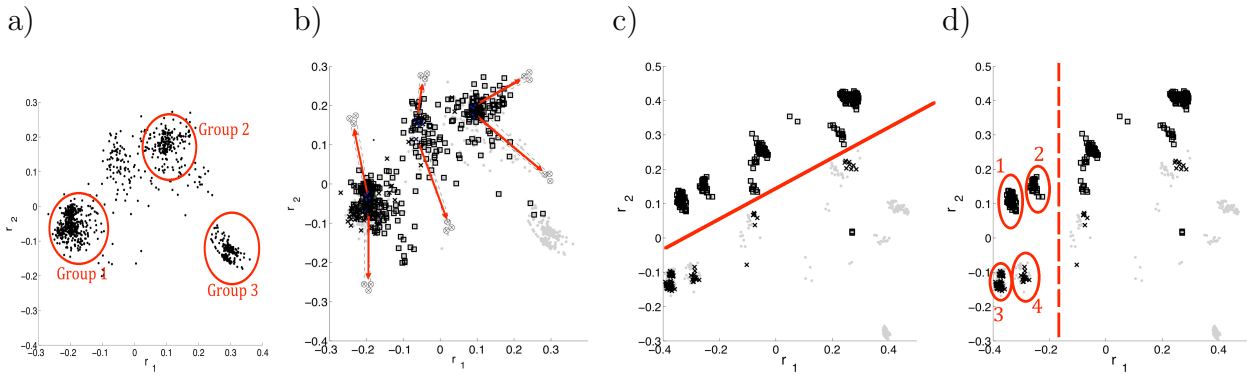


Figure 4: Figure a) provides an initial MDS view of the census data. The circles were added to draw attention to three clusters. Figure b) shows that students can mark observations based on ranges of *salary* ('×' and '□' show observations with salaries that are 'less than \$15k' and 'with than \$30k and \$60k', respectively) and drag observations (denoted by arrows) to inject feedback (as defined in Section 1) into visualizations. In response to the feedback, BaVA reconfigures the data. Figure c) displays the reconfiguration. We added a dotted line to show that the marked observations from Figure b) separate. In Figure d) we add circles to reference 4 clusters of interest.

display the data. Rather, we provide information on how to interpret and use Figure 4a to explore the data. In this case, each data point in Figure 4 represents an individual's completed survey. Although the axes of the visualization do not have an explicit physical meaning, the distance between any pair of surveys conveys the degree to which they are similar; e.g., surveys that appear in clusters are more similar to one another (according to the 68 data features) than surveys that appear in different clusters. However, the display, as currently plotted, does not convey *how* the surveys differ. That is, the mathematical method (MDS) used to create Figure 4a weighted the data features equally. Thus, to learn the features that differentiate the surveys, the students must explore the data and interact with the display; e.g., students may high-light observations according to requested criteria and/or change the perspective of the visualization by taking advantage of BaVA machinery.

For example, suppose that some students focus on the word "socio-economic" in the case study description and want to learn if there are features that correlate with the variable *salary*. Given the obvious structure in Figure 4a, these students might first identify three clusters and use high-lighting to discover that Group 1 represents surveys from working class people, Group 2 includes surveys from unemployed adults, and Group 3 includes surveys from adults under 20 years of age. Since none of the clusters are based purely on salary, the students may next highlight surveys based on two salary ranges: 'less than \$15k' or 'within \$30k and \$60k.' Figure 4b marks the surveys with the respective salary ranges by '$\times$' or '$\square$.' The marked observations do not present a clear clustering structure. This means that the display does not rely heavily on *salary* to differentiate observations. To change the perspective of the display and up-weight the role of *salary* in the display, the students may drag the marked observations from each group apart (the arrows in Fig-

MDS: A data visualization scheme that preserves pairwise distances in high-dimensional observations within a low-dimensional data representation (e.g., in two dimensions). Within the context of the census data from Section **??**, denote every high-dimensional observation i (for $i \in \{1, ..., n\}$) by $d_i = (d_{i,1}, ..., d_{i,p})$ ($p = 68$). MDS solves a two-dimensional, reduced version, $r_i = (r_{i,1}, r_{i,2})$, of each data point. The solution minimizes the difference in pairwise distances within $\boldsymbol{D} = [d_1, ..., d_n]'$ and $\boldsymbol{R} = [r_1, ..., r_n]'$; e.g., the distance $\|r_a - r_b\|$ between points r_a and r_b may approximate $\delta_{a,b}^{(\boldsymbol{d})}$, the distance between d_a and d_b. That is,

$$\boldsymbol{R} = \min_{r_1,...,r_n} \sum_{i<j\leq n} \left| \|r_i - r_j\| - \delta_{i,j} \right|, \qquad (1)$$

$$\delta_{i,j} = \sum_{d=1}^{p} Dist(d_{i,d} - d_{j,d}),$$

where $Dist(\cdot)$ represents a univariate distance function; e.g., euclidean distance. Solving Equation (1) is an optimization problem for which closed form expressions exist under certain mathematical constraints.

WMDS: WMDS is equivalent to MDS, but includes a $p-$vector of weights, $\boldsymbol{w} = \{w_1, ..., w_p\}$ (where, $\sum_d w_d = 1$) within the function $Dist(\cdot)$, e.g.,

$$\delta_{i,j} = \sum_{d=1}^{p} Dist(d_{i,d} - d_{j,d}) w_d$$

so that some dimensions in data $\boldsymbol{D}$ impact the solution for $\boldsymbol{R}$ more than others. When $w_i = w_j$ for all $\{i, j\} \in \{1, ..., p\}$, WMDS and MDS solve for equal values of $\boldsymbol{R}$.

ure 4b depict dragging). In turn, the software implements the BaVA process to reconfigure the visualization, as shown in Figure 4c. Now, the data appear in several small clusters and *salary*, in part, explains the spatialization of the clusters. We add a line to Figure 4c to show that the marked observations from Figure 4b separate perfectly; those above and below the line have surveys with salaries 'within \$30k and \$60k' and 'less than \$15k,' respectively.

Stopping the data exploration here would not support the students' goal to address the President's concerns and assess features correlated with *salary*. Thus, it is up to the students to assess which variables work jointly with *salary* to create the cluster structure in Figure 4c. One advantage of using the WMDS BaVA software is that, unlike Figure 4a, Figure 4c weights some data features more than others in response to the students' feedback in Figure 4b. The data features with the highest weights are the following: *Salary* (29%), *Have a reliable form of transportation to work* (20%), *Whether or not employed* (25%), and *Years of education* (10%). With this information, students may mark observations in Figure 4c to discover that 1) all observations for which $r_1 < -0.2$ represent *employed* individuals, 2) clusters 1 and 2 include individuals who make 'within \$30k and \$60k', but do or do not *have reliable modes of transportation to work*, 3) clusters 3 and 4 include individuals who make 'less than \$15k' and either 'drive themselves to work' or 'take public transportation' respectively. Now, students may conjecture that people with low-incomes need transportation assistance.

We expect students to make several conjectures about the data based on their visual explorations. The students report their findings in journals and, at the end of Phase I, during oral presentations. In the next phase, the students learn the mathematical and computational methods driving the visualization. An understanding of these methods may (or may not) impact their interpretations of the data.

EoT #1-5: The students assess their *points of view*, state the *goal*, and ask *questions*; gain an appreciation for the need of *information/data* to address questions; and *interpret* data visualizations to *infer* relationships in the data.

3.2 Phase II. Learn Mathematical Methods

Phase I does not require students to master mathematical concepts for data exploration. Now, in the second phase, the students learn the mathematical theory of MDS (as explained in Section 2 and in the sidebar), as well as its constraints. The students complete standard problems sets to reinforce the mathematical concepts. At the conclusion of the phase, students conjecture and formulate mathematically how displays based on MDS may change, given changes in its theory.

EoT # 5,6,7: The students learn the mathematical formulations of visualizations that rely on *assumptions* and result in *interpretations* which may lead to *inference*.

3.3 Phase III. Implement Computation

In Phase I, the students use software that implements the BaVA machinery based on the mathematics of Phase II. Now, the students program one or more modules within the software to re-implement BaVA. The software is coded in a way that includes self contained modules which, when removed, can be replaced by code created by students. By replacing modules, students are shielded from high-level coding. The modules that the students will replace include those that 1) read large high-dimensional datasets and 2) solve for coordinates $\boldsymbol{R}$ using a variety of techniques.

Since some students may not have computer programming in their backgrounds, computer lab assignments are important and Phase III may last longer than other phases. Note

that those experiencing programming for the first time have the benefit of a clear motivation to learn tedious (arguably), fundamental concepts, including, variable initialization, `if/then` statements, and loops.

EoT #5,6: Phase III reinforces the importance of summarizing and *interpreting* data using mathematical and computational *concepts* and *models*

3.4 Phase IV. Reflect

Now that the students have explored the data, learned the mathematics of MDS, and programmed it, the students have an opportunity to assess both the technical methods used to visualize the data and their personal thoughts while assessing and interpreting information in the data.

In regards to methods, the students experience in Phase I the need to adjust data displays, but only learn during Phases II and III a deterministic approach for summarizing data. Thus, in Phase IV, the students hypothesize, formulate mathematically, and implement how the visualization can adjust to their data interactions. Similar to Phase III, the students replace the BaVA software module that parameterizes feedback and adjusts data displays.

When students drag observations together or apart, the students are suggesting that the dimensions for which the observations are similar or different, respectively, are more important than the remaining dimensions; the weights of the important dimensions should be higher than the remaining weights. One way to incorporate this information is a data display is to apply Weighted MDS (WMDS). WMDS allows users to control a weight vector $\boldsymbol{w} = \{w_1, ..., w_p\}$ that reflects the degree to which each data dimension influences the visualization. WMDS results from a slight change in the MDS stress function. Students can solve for the weight by inverting the WMDS optimization and fixing the locations of the adjusted observations.

During Phase IV, the students also reflect upon what they gained from the data. During the reflection, they address the goals of the case study, state whether they validated their hypothesizes or corrected any misconceptions, and discuss any personal or analytical constraints. At the conclusion of Phase IV, students share their reflections and present their findings during an oral presentation and within a paper.

EoT #6,7,8,1: The students 1) evaluate the *model* and its *interpretation* given certain *assumptions* and 2) reflect upon *implications* (based on their *points of view*) of what they learned from the data and the role data served in making recommendations to the President.

Table 1: Lessons to teach MDS and critical thinking are motivated by real-world problem that includes the use of census data. The lessons group into four phases that cover one or more of the 8 Elements of Thinking (EoT). Also, during the phases, the students experience both the quantitative and the qualitative aspects of problem-solving.

Phase	**EoT**	**Quant/Qual**	**Description**
I. Assess/ Explore	1-5	Qualitative	Evaluate personal understanding and issues in the case study; State specific questions and hypotheses that might be learned from the data; Describe revelations from data exploration
		Quantitative	Learn the meaning of the visualization; Assess if it reflects personal judgement; Make adjustments, as needed
II. Methods	5, 6, 7	Qualitative	Hypothesize assumptions or weaknesses of WMDS
		Quantitative	Learn WMDS theory and math skills therein
III. Implemer	5, 6	Qualitative	Assess efficiency of code
		Quantitative	Gain computational skills
IV. Reflect	1, 7, 8	Qualitative	Assess WMDS and implementation; Evaluate weaknesses again; Design how to change theory and code to correct weaknesses; Assess personal biases; Reflect on what is learned; Suggest solutions or findings for the case study; Quantify gains of using data
		Quantitative	Implement theory and code adaptions

4 Discussion

Using DA (e.g., statistics) as a platform to emphasize critical thinking is not a new idea. However, the way by which we propose to integrate critical thinking with complex mathematical and computational methods is. We use IDVs so that students *construct* their understanding of 1) how to think critically, 2) the role of data in critical thinking, and 3) the mathematical and computational methods needed to summarize high-dimensional data. Additionally, as students progress through the phases, they have many opportunities to both assess what they understand and correct misconceptions.

Here, we only presented ideas within the context of MDS and WMDS. However, they can apply easily to other DA methods, including PCA and Classification and Regression Trees (CART). In time, we hope to develop an introductory DA course based on IDVs. Each unit in the course will correspond to a different DA method, rely on a different real-world case study, and apply the four phases. In this course, students will have multiple opportunities to experience the EoT and master not only DA methods, but also the thought process needed to solve problems with data.

References

Elder, L. and Paul, R. (2010), *A Thinker's Guide to Analytic Thinking*, Foundations of Critical Thinking, `http://www.criticalthinking.org/store-page.cfm?go=1&itemID=171&P=products&cateID=132&subcatID=0&catalogID=224`.

Endert, A., Han, C., Maiti, D., House, L., Leman, S., and North, C. (2011), "Observation-level Interaction with Statistical Models for Visual Analytics," Tech. rep., Virginia Tech.

House, L., Leman, S., and Han, C. (2010), "Bayesian Visual Analytics (BaVA)," Tech. Rep. 10-2, FODAVA, `http://fodava.gatech.edu/node/34`.

Jolliffe, I. (2002), *Principal Component Analysis*, John Wiley and Sons, Ltd, 2nd edn.

Kruskal, J. B. and Wish, M. (1978), "Multidimensional Scaling," *Sage University Paper series on Quantitative Application in the Social Sciences*, 48, 07–011.

Thomas, J. and Cook, K. (eds.) (2005), *Illuminating the Path*, National Visualizations and Analytics Center.

UCI, M. L. R. (1990), "US Census Data (1990) Data Set," `http://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)`.